

Relation between Gene Content and Taxonomy in Chloroplasts

Bashar Al-Nuaimi^{*†}, Christophe Guyeux^{*}, Bassam AlKindy[‡], Jean-François Couchot^{*}, and Michel Salomon^{*}

^{*}FEMTO-ST Institute, UMR 6174 CNRS, DISC Computer Science Department

Université de Bourgogne Franche-Comté, France

[†]Department of Computer Science, University of Diyala, Iraq

[‡]Department of Computer Science, University of Mustansiriyah, Iraq
christophe.guyeux@univ-fcomte.fr

Abstract The aim of this study is to investigate the relation that can be found between the phylogeny of a large set of complete chloroplast genomes, and the evolution of gene content inside these sequences. Core and pan genomes have been computed on *de novo* annotation of these 845 genomes, the former being used for producing well-supported phylogenetic tree while the latter provides information regarding the evolution of gene contents over time. It details too the specificity of some branches of the tree, when specificity is obtained on accessory genes. After having detailed the material and methods, we emphasize some remarkable relation between well-known events of the chloroplast history, like endosymbiosis, and the evolution of gene contents over the phylogenetic tree.

Index Terms—Chloroplasts, Phylogeny, Taxonomy, Core and Pan genomes, Gene content

I. INTRODUCTION

Understanding the evolution of DNA molecules is a very complex problem, and no concrete and well established solution exists at present regarding the case of large DNA sequences. Our objective in this article is to start to show that this complex problem can be (at least partially) solved when considering genomes of reasonable size and who faced a rational number of recombination, like in the chloroplasts case. However, various difficulties remain to circumvent when dealing with such a specific case, and solving them require the design of new ad hoc tools. Candidates for such tools are presented in this article, and are applied on the chloroplast case.

Chloroplasts are one of the numerous types of organelles in the plant cell. The term of chloroplast comes from the combination of chloro and plastid, meaning that it is an organelle found in plant cells that contains the chlorophyll. Chloroplast has the ability to convert water, light energy, and carbon dioxide (CO_2) into chemical energy by using carbon-fixation cycle [1] (also

TABLE I: Information on chloroplast sizes at highest taxonomic level

Taxonomy	nb. of genomes	min length	max length	average	standart deviation
Alveolata	4	85535	140426	115714.2	19648.3
Cryptophyta	2	121524	135854	128689.0	7165.0
Euglenozoa	7	80147	143171	98548.7	19784.5
Haptophyceae	3	95281	107461	102683.6	5307.6
Rhodophyta	9	149987	217694	183755.5	18092.2
Stramenopiles	35	89599	165809	124895.1	15138.0
Viridiplantae	775	80211	289394	150194.9	20376.8

called *Calven Cycle*, the whole process being called photosynthesis). This pivotal role explains why chloroplasts are at the basis of most trophic chains and are thus responsible for evolution and speciation.

Consequently, investigating the evolutionary history of chloroplasts is of great interest, and our long-term objective is to explore it by the mean of ancestral genomes reconstruction. This reconstruction will be achieved in order to discover how the molecules have evolved over time, at which rate, and to determine whether this way can present evidence of their cyanobacteria origin. This long-term objective necessitates numerous intermediate research advances. Among other things, it supposes to be able to apply the ancestral reconstruction on a well-supported phylogenetic tree of a representative collection of chloroplastic genomes. Indeed, sister relationship of two species must be clearly established before trying to reconstruct their ancestor. Additionally, it implies to be able to detect content evolution (modification of genomes like gene loss and gain) along this accurate tree. In other words, *gene content evolution* on the one hand, and *accurate phylogenetic inference* on the contrary, must be carefully regarded in the particular case of chloroplast sequences, as the two most important prerequisites in our quest of the last universal common ancestor of these chloroplasts.

The objective of this research work is to make significant progress in this quest, by providing material

TABLE II: Example of genomes information of *Streptophyta* clade

Organism name	Accession number	Sequence length	Nb of CDS
<i>Epimedium sagittatum</i>	NC_029428.1	158273	85
<i>Berberis bealei</i>	NC_022457.1	164792	267
<i>Torreya fargesii</i>	NC_029398.1	137075	100
<i>Lepidozamia peroffskyana</i>	NC_027513.1	165939	93
<i>Actinidia chinensis</i>	NC_026690.1	156346	271
<i>Quercus aliena</i>	NC_026790.1	160921	259
<i>Quercus aquifolioides</i>	NC_026913.1	160415	176
<i>Sedum sarmentosum</i>	NC_023085.1	150448	99

and methods required in the study of chloroplastic sequence evolution. Contributions of this article consist in the computation of core and pan genomes of the 845 complete genomes available on the NCBI, in the production of a well-supported phylogenetic tree based on core sequences as large as possible, and on the study of the produced data. In particular, we start to emphasize some links between the phylogenetic tree and evolution of gene content.

The paper is structured as follows. In the next section, material and methods applied in this study are presented, which encompass genome acquisition and annotation, core and pan genome analysis, and phylogenetic investigations. Obtained results related to such analyzes are detailed in Section III, on the chloroplast case. This article ends with a conclusion section, in which the study is summarized and intended future work is outlined.

II. MATERIALS AND METHODS

A. Data acquisition

A set of 845 chloroplastic genomes (green algae, red algae, gymnosperms, and so on) has been downloaded from the NCBI website, representing all the available complete genomes at the date of March, 2016 (see Table I). An example of such sequences, taken from the *Streptophyta* clade (a *Viridiplantae*), is provided in Table II. Note that this set does not really constitute a very balanced representation of the diversity of plants, as plants of particular and immediate interest to us like *Viridiplantae* are first sequenced. We must however deal with such bias, as genomic data acquisition is most of the time human-centred. This set of sequences presents too a certain variability in terms of length, as detailed in Table I.

Each genome has been annotated with DOGMA [2], an online automatic and accurate annotation tool of organellar genomes, following a same approach than in [3]. To apply it on our large scale, we have written (with the agreement of DOGMA authors) a script that automatic send requests to the website. By doing such annotations, the same gene prediction and naming process has been

TABLE III: Summarized properties of the pan genomes at the highest taxonomic level.

Taxonomy	Nb. genomes	Min N.b of pan genes	Max N.b of pan genes	Average Nb. of pan genes
<i>Alveolata</i>	4	253	266	262.25
<i>Cryptophyta</i>	2	258	259	258.5
<i>Euglenozoa</i>	7	193	267	253.428
<i>Haptophyceae</i>	3	251	266	258.333
<i>Rhodophyta</i>	9	156	267	246.222
<i>Stramenopiles</i>	35	73	271	238.971
<i>Viridiplantae</i>	775	85	271	229.827

applied with the same average quality of annotation. In particular, when a gene appears twice in the considered set of genomes, it receives twice the same name (no spelling error). At this level, each genome is then described by an ordered list of gene names, with possible duplications (other approaches for the annotation stage are possible, see, e.g., [3]). This description will allow us to investigate, later in this article, the evolution of gene content among the species tree, leading to the study of core and pan genomes recalled below.

B. Core and pan genome

Given a collection of genomes, it is possible to define their core genes as the common genes that are shared among all the species, while the pan genome is the union of all the genes that are in at least one genome (*all* the species have each core gene, while a pan gene is in *at least one* genome). Shared genes are evidences of evolution from a common ancestor and of the relatedness of chloroplast organisms.

To distinguish and determine the core genes may be of importance either to identify the specificity and the shared functionality of a given set of species, or to evaluate their phylogeny using the largest set of shared coding sequences. In the case of chloroplasts, an important category of genome modification is indeed the loss of functional genes, either because they become ineffective or due to transfer to the nucleus. Thereby, a small number of gene loss among species may indicate that these species are close to each other and belong to a similar lineage, while a significant loss means distant lineages.

So core genome is obviously of importance when inferring the phylogenetic relationship, while accessory genes of pan genome explain in some extend each species specificity. We have formerly proposed three approaches for eliciting core genomes. The first one uses correlations computed on predicted coding sequences [3], while the second one uses all the information provided during an accurate annotation stage [4]. The third method takes the advantages from the first two approaches, by considering gene information and DNA sequences, in order to find the targeted core genome [5].

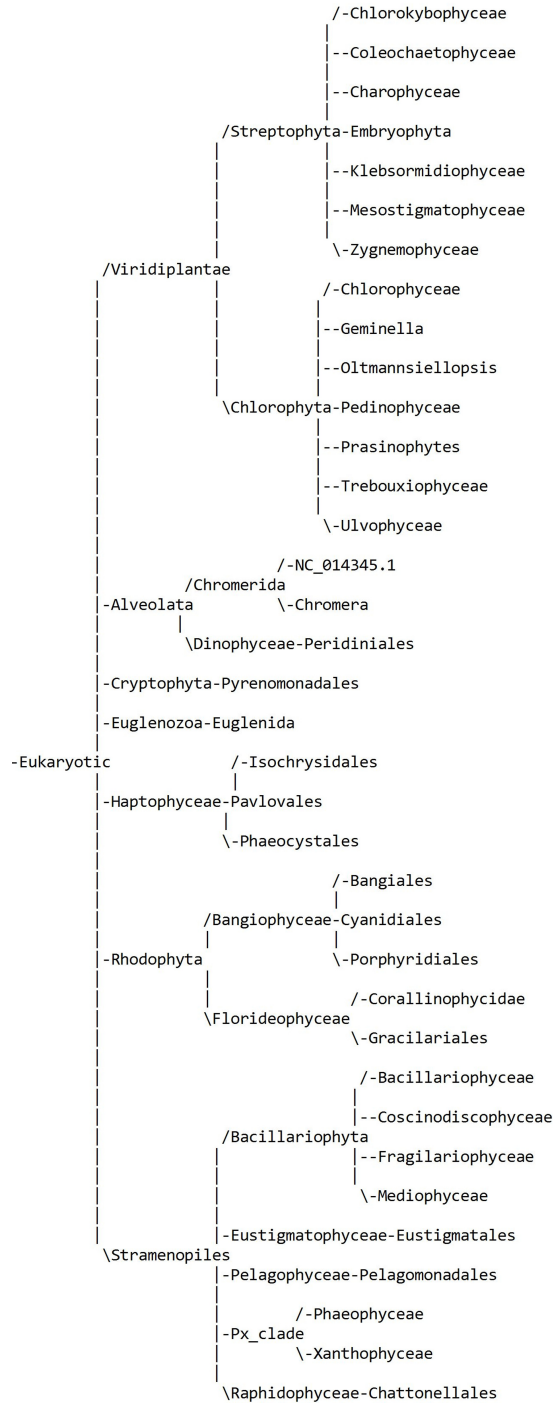


Fig. 1: Phylogenetic tree overview

We have found the core genome¹ of each selected family by using the second method described in the previous paragraph [4]. Obtained results regarding gene content are discussed in the next section. The core genome has been used too for our phylogenetic investigation

¹All data are available at...

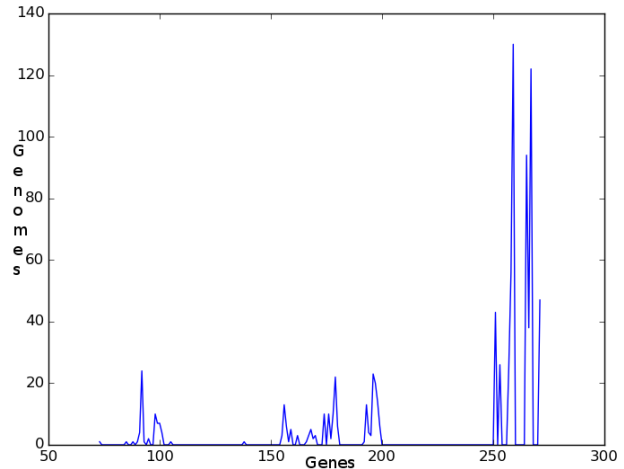


Fig. 2: The distributions of chloroplast genomes depending on the genomes size.

of chloroplast sequences, which has been applied as described hereafter.

C. Phylogeny study

The next step when trying to reconstruct the evolution of gene content over time is to deeply investigate the phylogeny of these chloroplasts, in order to obtain a tree as supported as possible. Indeed, a branching error in the tree may lead to an erroneous transmission of an ancestral state, which is dramatically perpetuated until reaching the last universal common ancestor. However, as we considered all existing plant taxa, we faced chloroplastic sequences that have diverged a lot since two billion of years, so the core genome of these 845 sequences is very small when compared with sequence length of each representative, and inferring a tree on such a partial information will probably lead to numerous errors.

The approach that has been regarded in our study was then to group the plant families per close packets (same family in the taxonomy). Such grouping has enlarged the number of shared gene sequences (core genes of the considered family) on which a more representative phylogeny can be computed [6]. After having aligned the core genes of each family using MUSCLE [7] on our supercomputer facilities, we then have inferred a phylogenetic tree per family.

To obtain such a tree, the RAxML [8], [9] program has been employed to compute the phylogenetic maximum-likelihood (ML) function with the setup described hereafter. General Time Reversible model of nucleotide substitution, with Γ model of rate heterogeneity and hill-climbing optimization method. The *Prochlorococcus marinus* (NC_009091.1) cyanobacteria species has

finally been chosen as outgroup, due to the supposed cyanobacteria origin of chloroplasts.

After such a computing, if all bootstrap values are larger than 95%, then we have considered that the phylogeny is resolved, as the largest possible number of genes has led to a very well supported tree. In case where some branches are not supported, we can wonder whether a few genes can be incriminated in this lack of assistance, for a large variety of reasons encompassing homoplasy, stochastic errors, undetected paralogy, incomplete lineage sorting, horizontal gene transfers, or even hybridization. Such problem has been resolved by finding the largest subset of core genes leading to the most supported tree, by the heuristic approach coupled with statistical LASSO tests described in [6], [10]. Obtained trees are then merged on a well-supported and representative supertree.

III. OBTAINED RESULTS

A. Phylogenetic investigations

The approach detailed in the previous section has led to a well supported phylogenetic tree of the whole available chloroplasts, with the ordered list of genes at each leaf of the tree. An overview of the latter is provided in Figure 1. Obtained tree available too on our website is in general coherent with the NCBI taxonomy, except in some specific locations.

By going into the details of the obtained tree, it is well known that the first plants endosymbiosis ended in a great diversification of lineages comprising Red Algae, Green Algae, and Land Plants (terrestrial). The interesting point in the production of our results is that the organisms resulting from the first endosymbiosis are distributed in each of the lineages found in the chloroplast genome structure evolution as outlined in Figure 1.

More precisely, all Red Algae chloroplasts are grouped together in one lineage, while Green Algae and Land Plant chloroplasts are all in a second lineage. Furthermore, organisms resulting from the secondary endosymbioses, as listed in Table IV, are well localized in the tree: both the chloroplasts of Brown Algae and *Dinoflagellates* representatives are found exclusively in the lineage also comprising the Red Algae chloroplasts from which they evolved, while the *Euglens* is related to Green Algae from which they evolved. This latter makes sense regarding biology, history of lineages, and theories of chloroplasts origins (and so photosynthetic ability) in different *Eucaryotic* lineages [11].

B. Gene content

Let us now investigate the gene content level of the tree. Indeed, genes are rearranged in the genome

by evolutionary events like insertion, deletion, transposition, and inversion, which are called genome rearrangements [12]. Such rearrangements can be studied, considering that we have both the gene contents and the phylogeny. A general overview of obtained results, in terms of gene contents (pan genome) evolution at the top taxonomy level, is provided in Table III, and it is detailed for the following taxonomic level in Table IV.

The core genome is constituted by 36 coding sequences, namely: *ATPA*, *ATPB*, *ATPH*, *ATPI*, *PETB*, *PETG*, *PSAA*, *PSAB*, *PSAC*, *PSAJ*, *PSBA*, *PSBC*, *PSBD*, *PSBE*, *PSBF*, *PSBH*, *PSBI*, *PSBJ*, *PSBL*, *PSBN*, *PSBT*, *PSI_PSBT*, *RBCL*, *RPL14*, *RPL16*, *RPL2*, *RPL20*, *RPL36*, *RPS11*, *RPS12*, *RPS12_3END*, *RPS14*, *RPS19*, *RPS2*, *RPS7*, and *RRN16*. The pan genome of the whole considered species, for its part, contains 268 genes. Note that, according to our computation, no gene was specific to a given clade (that is, present in only one clade).

C. Relations between gene content and phylogeny

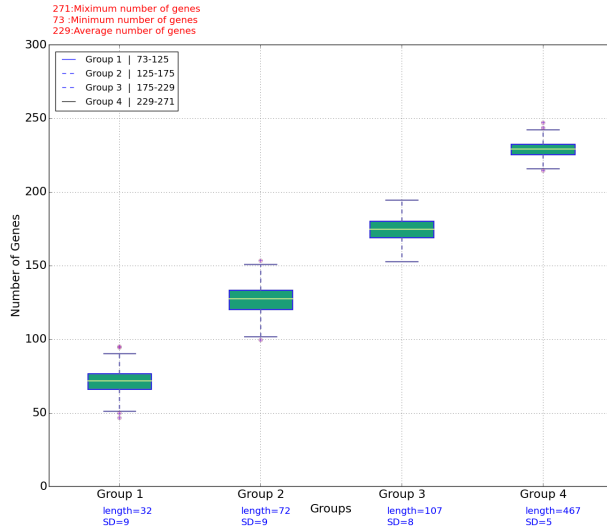
We then have further investigated the distribution of number of genes according to the group of species. Obtained results are reproduced in Figures 2 and 3. Four groups have appeared among the 845 genomes, which are taxonomically coherent. As shown in Fig. 3, the cluster of largest genomes has a number of genes ranging from 229 to 271, while in the group of smallest genomes, the lowest number of genes is for the *Viridiplantae* case. In particular, among the genomes having less than 120 genes, we found accession number NC_012903.1 (*Eukaryota*, *Stramenopiles*, *Pelagophyceae*, *Pelagomonadales*, *Aureoumbra lagunensis*), and 63 *Spermatophyta* species: 3 *Pinidae*, 58 *Magnoliophyta*, one *Cycadidae*, and finally one *Gnetidae*. We finally obtain chloroplast genomes varying from 73 to 271 genes.

We can further note that (1) most of the organisms in green lineage (green algae and land plants) have a lower number of genes in their chloroplasts compared to the red algae. (2) Most land plants have genome sizes ranging between 120 and 160 kb [13]. (3) Most of the differences in genome size are due to the number of paralogous genes.

When regarding more deeply the ordered list of genes to investigate the reasons of such differences of size, it appears to us that the gene content evolution can mostly be explained by repetitions of some genes and the loss of other ones: no large scale recombination is responsible of such variations. Usual case is as in Figure 4 for ACCA pan gene, on which single vulnerable genes are lost, possibly in various independent branches, due to deleterious mutations. Such results have been obtained by comparing, for each couple of close genomes, all gene names and positions, by practicing a naked eye investigation using homemade scripts. Some mutation

TABLE IV: Taxonomy in the second level

Taxonomy		Nb. genomes	Min N.b of pan genes	Max N.b of pan genes	Avg N.b of pan genes
<i>Alveolata</i>	<i>Chromerida</i>	2	253	265	259.0
	<i>Dinophyceae</i>	2	265	266	265.5
<i>Cryptophyta</i>	<i>Pyrenomonadales</i>	2	258	259	258.5
<i>Euglenozoa</i>	<i>Euglenida</i>	7	193	267	253.428
<i>Haptophyceae</i>	<i>Phaeocystales</i>	1	266	266	266.0
	<i>Isochrysidales</i>	1	251	251	251.0
	<i>Pavlova</i>	1	258	258	258.0
<i>Rhodophyta</i>	<i>Bangiophyceae</i>	6	156	266	240.166
	<i>Florideophyceae</i>	3	251	267	258.333
<i>Stramenopiles</i>	<i>Px_clade</i>	6	251	271	261.166
	<i>Bacillariophyta</i>	20	138	271	231.35
	<i>Eustigmatophyceae</i>	6	253	267	262.16
	<i>Raphidophyceae</i>	1	258	258	258.0
	<i>Pelagophyceae</i>	2	73	266	169.5
<i>Viridiplantae</i>	<i>Chlorophyta</i>	58	156	271	244.517
	<i>Streptophyta</i>	717	85	271	228.638

**Fig. 3:** Classification of chloroplast genomes according to numbers of pan genes.

and indel events are provided too in Table V, for the sake of illustration.

IV. CONCLUSION

In this article, we made significant progress in the study of chloroplastic sequence evolution, by providing material and methods required in the quest of the ancestral genome of the chloroplasts. A large set of complete chloroplast genomes has been studied *de novo* regarding both core and pan genomes, phylogenetic relationship, and gene content modifications. We then started to study the produced data, by emphasizing some remarkable relations between well-known events of the chloroplast history and the evolution of gene contents over the phylogenetic tree.

In future work, our intention is to investigate more systematically such relations between remarkable ancestral nodes in the tree, endosymbiosis events, and evolution of gene content. We will wonder whether some branches of the trees are statistically remarkable when considering gene content (for instance, do we have a correlation between the presence or absence of a subset of genes, and a particular taxonomy). Then, the gene ordering and content of each ancestral node will be computed using ad hoc algorithms, ancestral DNA sequences will be inferred, and ancestral intergenic regions will be deduced, in order to have all ancestral genomes with confidence indications like probabilities. The produced ancestral genomes will then be used to investigate hypotheses formulated by biologists, regarding the origin of chloroplasts, their recombination events, and the transfer of some material to the nucleus. We will in particular study whether recombination events were uniform over time and on the whole sequence, or if it is possible to highlight some hot spots of recombination in the history of these chloroplasts.

All computations have been performed using the “Méso-centre de calcul de l’Université de Franche-Comté” supercomputer facilities.

REFERENCES

- [1] J. Lewis M. Raff K. Roberts B. Alberts, A. Johnson, John H. Wilson P. Walter, and Tim Hunt. Molecular biology of the cell. *Biochemistry and molecular biology education*, 31(3):212–213, 2003.
- [2] Stacia K. Wyman, Robert K. Jansen, and Jeffrey L. Boore. Automatic annotation of organellar genomes with dogma. *BIOINFORMATICS, Oxford Press*, 20(172004):3252–3255, 2004.
- [3] Bassam Alkindy, Jean-François Couchot, Christophe Guyeux, Arnaud Mouly, Michel Salomon, and Jacques M. Bahi. Finding the core-genes of chloroplasts. *Journal of Bioscience, Biochemistry, and Bioinformatics*, 4(5):357–364, 2014.

TABLE V: Example of comparison between pairwise genomes from various species, to investigate the changes that occurred within branches of the tree.

Index	Clade	Sub-kingdom	Order/ Family	Genome name	N.b of pan genes	Deletion/ Insertion	Matching ratio
1	Viridiplantae	Embryophyta	Camelineae	Camelina	92	173/0	48.46
				Barbarea	267		
2	Viridiplantae	Embryophyta	Camelineae	Aquilaria	101	164/0	28.26
				Hibiscus	267		
3	Viridiplantae	Embryophyta	Sapindales	Acer	92	173/0	49.02
				Azadirachta	267		
4	Viridiplantae	Embryophyta	Zamiaceae	Lepidozamia	92	172/0	51.66
				Zamia	267		
5	Viridiplantae	Embryophyta	Lamiaceae	Lavanduleae	92	173/0	49.91
				Perman	267		
6	Stramenopiles	Pelagophyceae	Pelagomonadales	Aureoumbra	73	193/0	27.13
				Aureococcus	267		
7	Viridiplantae	Eudicotyledons	Berberidoideae	Epimedium	85	180/0	33.52
				berberis	267		
8	Viridiplantae	Eudicotyledons	Actinidia	NC_026690_1	92	174/0	48.63
				NC_026691_1	271		

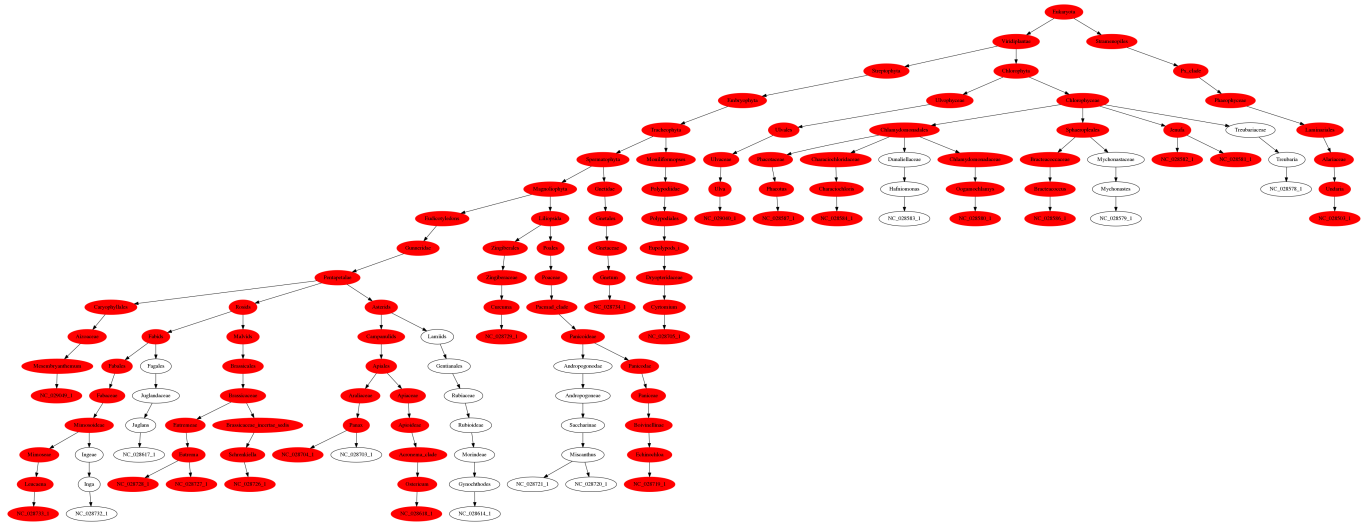


Fig. 4: ACCA gene loss in various branches of the tree

- [4] Bassam AlKindy, Christophe Guyeux, Jean-François Couchot, Michel Salomon, and Jacques M Bahi. Gene similarity-based approaches for determining core-genes of chloroplasts. In *Bioinformatics and Biomedicine (BIBM), 2014 IEEE International Conference on*, pages 71–74. IEEE, 2014.
- [5] Bassam AlKindy, Huda Al-Nayyef, Christophe Guyeux, Jean-François Couchot, Michel Salomon, and Jacques M. Bahi. Improved core genes prediction for constructing well-supported phylogenetic trees in large sets of plant species. In Francisco Ortuño and Ignacio Rojas, editors, *Bioinformatics and Biomedical Engineering*, volume 9043 of *Lecture Notes in Computer Science*, pages 379–390. Springer International Publishing, 2015.
- [6] Bassam AlKindy, Christophe Guyeux, Jean-François Couchot, Michel Salomon, Christian Parisod, and Jacques M. Bahi. Hybrid genetic algorithm and lasso test approach for inferring well supported phylogenetic trees based on subsets of chloroplastic core genes. *CoRR*, abs/1504.05095, 2015.
- [7] Robert C Edgar. Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, 32(5):1792–1797, 2004.
- [8] Alexandros Stamatakis. Raxml version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, page btu033, 2014.
- [9] Jeffrey Rizzo and Eric C Rouchka. Review of phylogenetic tree construction. *University of Louisville Bioinformatics Laboratory Technical Report Series*, pages 2–7, 2007.
- [10] Reem Alsrarj, Bassam AlKindy, Christophe Guyeux, Laurent Philippe, and Jean-François Couchot. Well-supported phylogenies using largest subsets of core-genes by discrete particle swarm optimization. *Proceedings of CIBB*, 2:1, 2015.
- [11] Xi Li, Ti-Cao Zhang, Qin Qiao, Zhumei Ren, Jiayuan Zhao, Takahiro Yonezawa, Masami Hasegawa, M James C Crabbe, Jianqiang Li, and Yang Zhong. Complete chloroplast genome sequence of holoparasite cistanche deserticola (orobanchaceae) reveals gene loss and horizontal gene transfer from its host haloxylon ammodendron (chenopodiaceae). *PloS one*, 8(3):e58747, 2013.
- [12] Mikita Suyama and Peer Bork. Evolution of prokaryotic gene order: genome rearrangements in closely related species. *Trends in Genetics*, 17(1):10–13, 2001.
- [13] Beverley R Green. Chloroplast genomes of photosynthetic eukaryotes. *The plant journal*, 66(1):34–44, 2011.